Supplementary Material

This material provides supplemental tables and figures that were referred in the main manuscript titled *Café and Restaurant Under My Home: Predicting Urban Commercialization Through Machine Learning.*

Supplemental Table S1. Conceptual Elements of Confusion Matrix

		Actual		Sum of predicted
		True	False	
Predicted	Positive	True positive (TP)	False positive (FP)	Predicted positive (TP + FP)
	Negative	False negative (FN)	True negative (TN)	Predicted negative (FN + TN)
Sum of actual		Actual true (TP + FN)	Actual false (FP + TN)	N (TP + FP + FN + TN)

Notes: This conceptual matrix shows relation between four different measures (accuracy, sensitivity, specificity, and balanced accuracy) that are used to evaluate machine learning model performance.

Supplemental Table S2. Weights and Booster Parameters in XGBoost Model

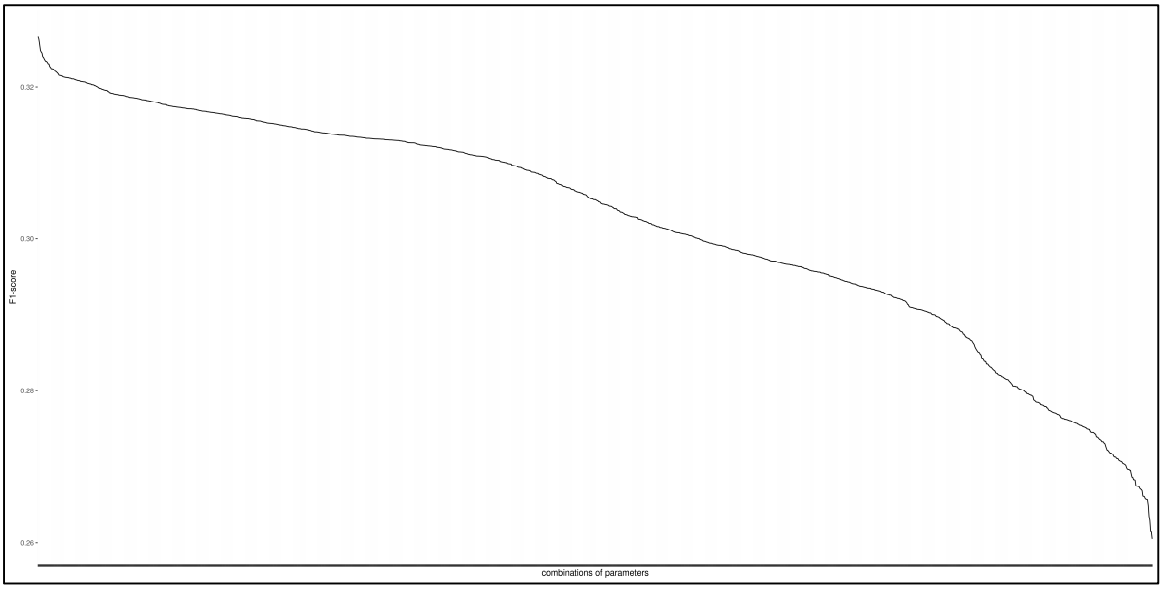
	Name	Range (Unit)	Final Value	Default Value
Weight				
Weight	scale_pos_weight	47.44	6.89	1
Square root weight	scale_pos_weight_sqrt	6.89		
Booster Parameter				
Maximum depth of tree	max_depth	3 to 10 (1)	3	6
Minimum sum of instance weight	min_child_weight	0 to 1 (0.1)	1	1
Minimum loss reduction	gamma	0 to 1 (0.1)	0.8	0
Subsample ratio of columns	colsample_bytree	0.5 to 1 (0.1)	0.9	1
Subsample ratio of training instances	subsample	0.4 to 1 (0.1)	1	1
L1 regularization term on weights	alpha	0 to 0.2 (0.02)	0.44	0
Learning rate	eta	0.1 to 1 (0.1)	0.3	0.3

Notes: Parameter name is based on xgboost version of 0.90.0.2. objective = "binary:logistic". booster = "gbtree". eval_metric = “aucpr".

Supplemental Figure S1. F1-score for Combinations of Parameters and Weights in XGBoost Model

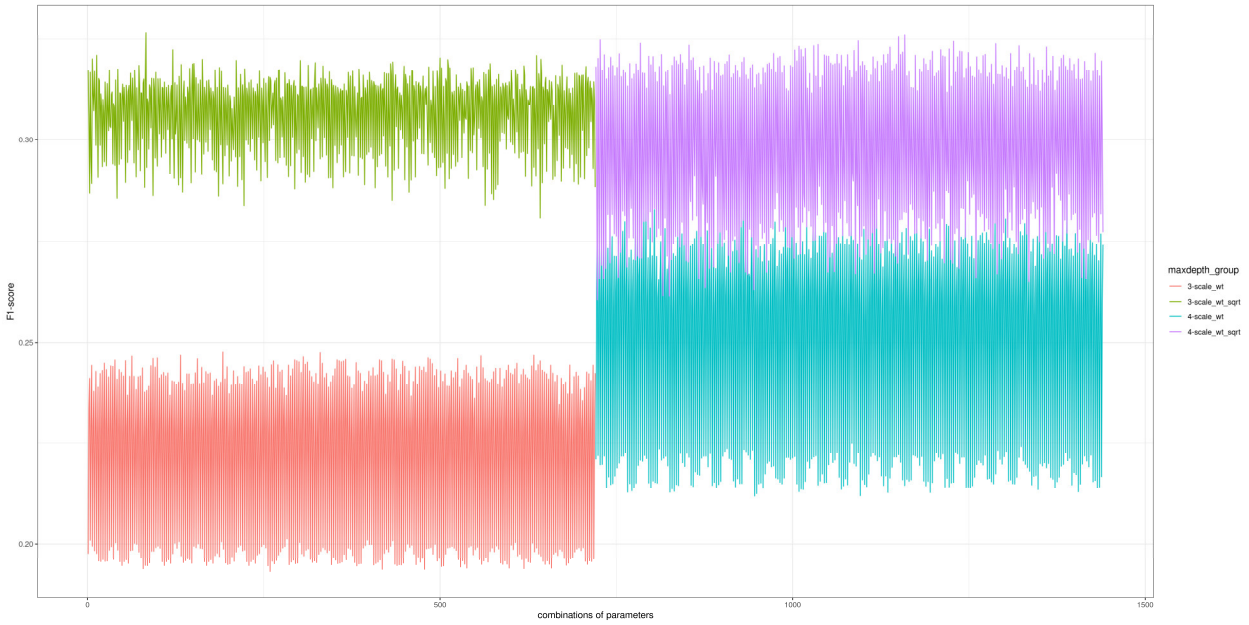
(a) Weight of 47.436

(b) Weight of 6.887



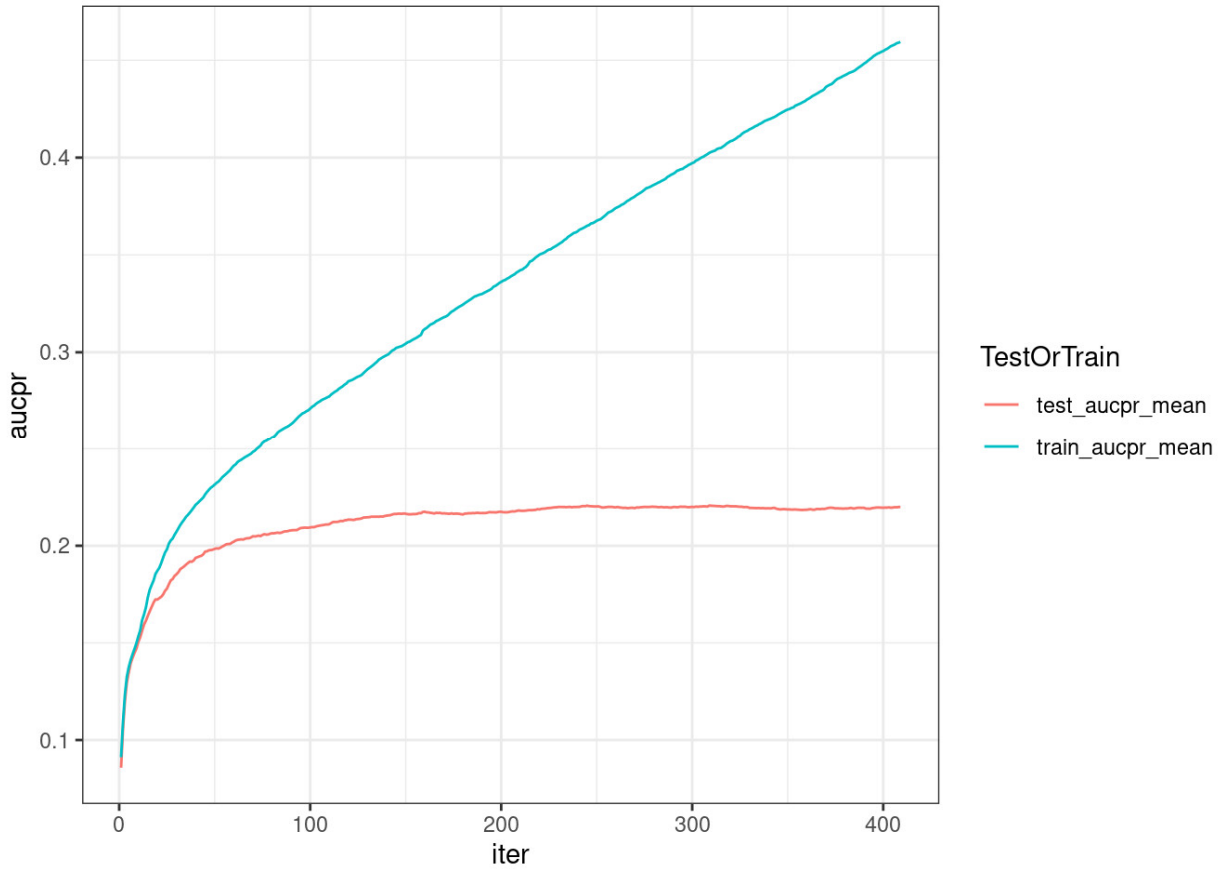
Notes: Y-axis represents F1-score and x-axis represents 8,131,200 combinations of parameters and weights. The combinations are ordered on x-axis to place the combination with the highest F1-score closest to y-axis and the combination with the lowest F1-score away from y-axis.

Supplemental Figure S2. F1-score for Cross-validation between Weight and Booster Parameter on the Maximum Depth of Tree



Notes: Y-axis represents F1-score and x-axis represents possible combinations of parameters and weights. The combinations are ordered on x-axis to place combinations with the booster parameter (maximum depth of tree) of 3 closest to y-axis and combinations with the parameter of 4 away from y-axis. Red highlights represent combinations with the weight of 47.436 and the parameter of 3, green is with the weight of 6.887 and the parameter of 3, blue is with the weight of 47.436 and the parameter of 4, and purple is with the weight of 6.887 and the parameter of 4.

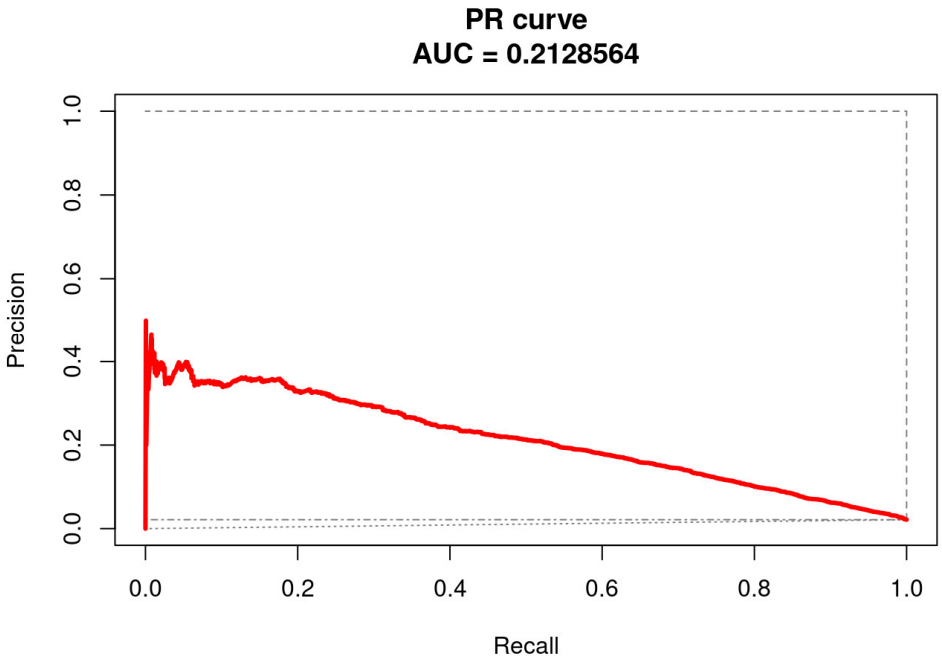
Supplemental Figure S3. Test Error and Train Error in XGBoost Model, By the Number of Iteration



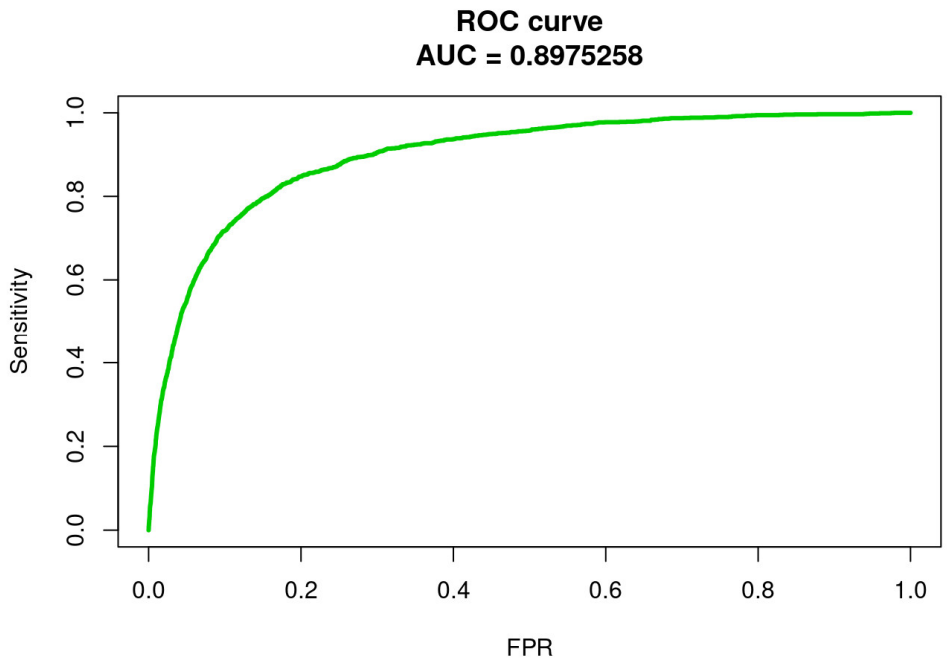
Notes: AUCPR on y-axis means the area under the precision-recall curve. The number of iteration on x-axis is displayed in this figure from 0 to 410 though its maximum is 1,000.

Supplemental Figure S4. Precision and Sensitivity of XGBoost Model

(a) AUC of PR curve



(b) AUC of ROC curve



Notes: AUC means area under curve; PR means precision recall; ROC means receiver operating characteristic. AUC of PR curve (0.213) and AUC of ROC curve (0.898) match Table 2 in the main manuscript.